

DOCUMENT RESUME

ED 463 308

TM 033 740

AUTHOR Kroff, Michael W.
TITLE Mutilating Data and Discarding Variance: The Dangers of Dichotomizing Continuous Variables.
PUB DATE 2002-02-14
NOTE 24p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (25th, Austin, TX, February 14-16, 2002).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Analysis of Variance; Heuristics; *Regression (Statistics); Reliability
IDENTIFIERS *Dichotomous Variables

ABSTRACT

This paper reviews issues involved in converting continuous variables to nominal variables to be used in the OVA techniques. The literature dealing with the dangers of dichotomizing continuous variables is reviewed. First, the assumptions invoked by OVA analyses are reviewed in addition to concerns regarding the loss of variance and a reduction in score reliability that result from the conversion of continuous variables to nominal variables. Second, regression is discussed as a more adequate alternative to analysis of variance (ANOVA). Finally, a heuristic data set is presented and used as a demonstration of three alternatives that can occur through this conversion and when ANOVA is used. (Contains 6 figures and 20 references.) (Author/SLD)

Running head: DICHOTOMIZING CONTINUOUS VARIABLES

ED 463 308

M. Kroff

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

✓ This document has been reproduced as
received from the person or organization
originating it.

12 Microfilm editions have been made to
improve reproduction quality.

• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

Mutilating Data and Discarding Variance: The Dangers of Dichotomizing Continuous Variables

Michael W. Kroff

Texas A&M University

TM033740

Paper presented at the annual meeting of the Southwest Educational Research
Association, Austin, TX, February 14, 2002.

Abstract

The present paper reviews issues involved in converting continuous variables to nominal variables to be used in OVA techniques. The literature dealing with the dangers of dichotomizing continuous variables is reviewed. First, the assumptions invoked by OVA analyses are reviewed in addition to concerns regarding the loss of variance and a reduction in score reliability that result from the conversion of continuous variables to nominal variables. Second, regression is discussed as a more adequate alternative to ANOVA. Finally, a heuristic data set is presented and used as a demonstration of three alternatives that can occur through this conversion and when ANOVA is employed.

Mutilating Data and Discarding Variance: The Dangers of Dichotomizing Continuous Variables

OVA techniques were historically the dominant method used in the social sciences (Edgington, 1964, 1974). Thompson (1985, 1986, 1994, 1999) has argued, however, that regression and other GLM techniques have recently become more prevalent while the proportion of OVA methods has declined (cf. Elmore & Woehike, 1988; Goodwin & Goodwin, 1985; Stone-Romero, Weaver, and Glenar, 1995). One reason for this shift is offered by Thompson (1999):

Improved analytic choices have partially been a function of growing researcher awareness that The researcher's fundamental task in deriving defensible results is to employ an analytic model that matches the researcher's (too often implicit) model of reality. (p. 24)

Thompson (1986) offered three potential weaknesses oftentimes involved in ANOVA. First, the requirement that continuous data be converted to nominal data may result in information being "squandered." Most variables have scales that are higher than nominal, a notable exception being gender. The reality of these variables (as continuous) is lost as they are modified (as nominal) for analytic purposes. The result is an analytic model that no longer matches the model of reality.

Second, the relationships and distribution shapes of some variables may become distorted through OVA methods (Thompson, 1986). The "balanced design," preferred by researchers employing OVA methods, is not typically indicative of reality. While having equal number of participants in each cell is advantageous for OVA analysis because effects are then uncorrelated, a high price is paid, as explained by Thompson (1986), "as

subjects rarely come or stay packaged in equal or proportional numbers per cell, researchers may induce further mortality in order to achieve balanced design" (p. 919). Furthermore, predictors are assumed to be normally distributed and perfectly uncorrelated in OVA analyses. In reality, however, the predictors may be normally distributed and may be somewhat correlated with each other (Thompson, 1986).

Third, the reliability of the predictors (that were originally continuous) decreases as they are converted to a nominal scale. Reliability is basically a function of variance. It follows that discarding information and its subsequent loss of variance (through conversion from continuous to nominal data) will result in a decrease in reliability. The result is a reduction in power against Type II errors (Cohen, 1968, Thompson, 1986, Thompson, 1999). A summation of problems inherent in ANOVA is offered by Humphreys (1978):

To summarize the defects in the analysis of variance design, primarily there is a substantial loss in the power of the statistical test of the research hypothesis.

There is often, as well, an illusion of control of variables and a tendency to interpret significant relationships causally. Also, when orthogonality is imposed

The purpose of this paper is to review the effects of converting continuous variables into ones that are dichotomized (or trichotomized). First, the nature of dichotomized variables is discussed. Next, the inherent problems that result from converting continuous variables to dichotomous ones are reviewed. Next, regression is discussed as an alternative to ANOVA. Finally, heuristic examples are offered that demonstrate these effects and how they alter the true nature of the original data.

Dichotomous Variables

A dichotomous variable is "a discrete variable with only two classifications" (Hinkle, Wiersma, & Jurs, 1998, p. 617). It is discrete in that the number of possible classifications can be counted (in this case- two). A dichotomous variable is one example of a categorical variable. A categorical variable is "one in which subjects differ in type or kind" (Kerlinger & Pedhazur, 1973, p. 102). Examples of categorical variables include race, biological sex, and marital status. Biological sex is also an example of a dichotomous variable in that there are exactly two classifications.

A continuous variable, on the other hand, is one "that can take on any value in the measurement scale being used" (Hinkle, Wiersma, & Jurs, 1998, p. 617). Subjects differ on continuous variables in amount or degree (Kerlinger & Pedhazur, 1973). Examples of continuous variables include intelligence, dosages of a drug, and frequency of reinforcement.

To dichotomize (or trichotomize) a continuous variable, which is required for any OVA analysis, one must divide the scale or results of the continuous variable into a fixed number of classifications (two for a dichotomy and three for a trichotomy). An example would be to take scores from an intelligence test and divide them into two or three groups. In this manner, the original scores take on new meaning as part of a new group with "similar" scores. Each group represents a level in a particular "way" for the purposes of ANOVA. In this case, a number of levels (determined by the number of groups created by the researcher) of the way- intelligence- are created.

A major problem inherent in converting continuous data to nominal (categorical) data is the loss of information (Kerlinger, 1986; Pedhazur, 1982). As Thompson (1981) emphasized:

When we reduce interval level of scale data to the nominal level of scale we are doing nothing less than thoughtlessly throwing away information which we previously went to some trouble to collect. If research is conducted for the purpose of acquiring knowledge, then is it consistent with our purpose to employ a method which "throws away" information which might provide a more refined understanding of the phenomena which we are studying? (p. 8)

This "throwing away of information" occurs in the case of dichotomized or trichotomized variables that were originally continuous because of two basic assumptions of dichotomized data. As explained by Thompson (1999), these assumptions are:

- 1- All participants in a given level of a way are considered to be the same.
- 2- All participants in a given level of a specific way are considered to be different from those in different levels of that way.

For example, if a researcher has a somewhat normal distribution of cholesterol scores and trichotomizes them at 125 and 175, he/she is saying that a person with a score of 130 is the same as a person with a score of 165. In addition, a person with a score of 173 is different than a person with a score of 177. Essentially, scores on a continuous variable (cholesterol scores) have been reduced to three levels of a trichotomized way (cholesterol way). A given person's cholesterol score, therefore, can belong to only one level and is equally different from all of the scores in the other two levels. This analytic

model created by converting the continuous data to nominal data is seemingly not indicative of reality.

Another major problem of dichotomizing variables is a reduction of reliability. Furthermore, as described by (Thompson, 1999):

In such cases, it is the reliability of the dichotomy that we actually analyze, and not the reliability of the highly-reliable, intervally-scaled data that we originally collected, which impact the analysis we are conducting. (p. 27, emphasis in original)

In other words, the reduction of a reliable variable to a dichotomous variable actually makes that variable *less* reliable (Cliff, 1987; Humphreys, 1978). This reduction results in less power against Type II error (Cohen, 1968; Thompson, 1986). The reduction in reliability occurs as emphasis shifts from the reliable data itself to how reliably the data was dichotomized or trichotomized. In the above example the researcher would no longer be analyzing the reliable scores of 130, 165, etc. but would instead be analyzing the new scores created by the trichotomized groups. In this example, the three groups (levels) would consist of those scores less than 125, those between 125 and 175, and those greater than 175.

As a heuristic example, consider a study by Song and Parry (1999) regarding product innovativeness and its relationship with various organizational synergies and product performance. In this study, Song and Parry collected data on 788 products from 404 firms. These firms responded to 7 measures of product innovativeness on scales of 0 to 10 (10 indicative of "very innovative"). The results are reproduced in Table 1.

As seen in Table 1, the majority of products (561) were rated between 2.5 and 7.5 on the averaged scale of innovativeness. Only 32 were rated between 0 and 2.5 while 195 were rated between 7.5 and 10. At this point the authors decided to dichotomize the results into "low-innovativeness" and "high-innovativeness" making the split at the score mean of 5.99. By making this split, the authors essentially forced the assumption that all scores between 0 and 5.99 are the same and that all scores between 6.0 and 10.0 are also the same. This seems hardly the case given the initial breakdown of results in Table 1.

This dichotimization has also forced the assumption that all scores in the "low-innovativeness" category are equally different from those in the "high-innovativeness" category. In other words, the difference between a score of 5 and 6 is the same as the difference between 2 and 9. Again, this does not seem to match the reality of the data.

Parry and Song (1999) continued their analysis by running an ANOVA with the two levels of innovativeness against six organizational and performance variables. A seemingly better alternative would have been to run a regression for each of the six dependent variables with innovativeness retained as a continuous independent variable. The results would then have been based on the reliability of the original data and not on the reliability of the dichotomy, which in this case, was seemingly less.

Regression can serve as a replacement for ANOVA because, as noted by Cohen (1968), ANOVA and ANCOVA are special cases of multiple regression. Furthermore, as pointed out emphatically by Cohen and Cohen (1983), "it (multiple regression/correlation analysis) is a versatile, all-purpose system for analyzing the data of the behavioral, social, and biological sciences and technologies" (p. 4, emphasis added).

Kerlinger and Pedhazur (1973) added:

... (multiple regression) can be used equally well in experimental or non-experimental research. It can handle continuous and categorical variables. It can handle two, three, four, or more independent variables... Finally... multiple regression analysis can do anything the analysis of variance does- sum of squares, mean squares, F ratios- and more. (p. 3, emphasis added)

Three Heuristic Scenarios

Whether dichotomizing or trichotomizing intervally-scaled data is appropriate depends on the situation. Specifically, it is dependent on whether the converted data meets the two assumptions of the new model of reality: First, is each member of a particular group essentially the same, and second, are all members of one group essentially different from the members of the other groups?

There are three possible situations that can influence the decision to convert intervally-scaled data to nominally-scaled data. These three alternatives are demonstrated using the heuristic data in Table 2.

Alternative one- Impact is minimal. In this scenario, the data in Table 2 for X1 are re-expressed as a trichotomy under X1'. This does not appear problematic in this case as the grouping of scores maintains the general variability of the original data set. In other words, the assumptions invoked by the trichotomizing of the data are consistent with original data. The scores within each group are basically the same, and the scores within each group differ equally from the scores in the other groups. The analytic model set up by the researcher matches the model of reality in the original scores. Graphically, this can be seen in Figures 1 and 2. In this case, the variance in the original data is

maintained in the trichotomy, and the results, therefore, will be representative of the original data.

Alternative two- Variance is created where there is none. This situation occurs when intervally-scaled data are converted without clear differences between the original scores. The data in X2 found in Table 2, for example, are clearly lacking variance between the scores. The conversion of this data however, as shown in X2', would represent an analytical model purporting a difference between groups that does not match the model of reality in which the difference is meaningless.

As evidence of the creation of imaginary variance, Thompson (1994) refers to a dissertation in which the author assigned children to one of three depression groups based on their scores on a depression variable. The top score of all of the children was 3.43. As noted by the author of the dissertation, the recommended *lower-end* cutoff for severely depressed people is 4. In other words, the author created three groups of depressed children when none of them met the minimum score for being clinically depressed!

Alternative three- Variance is discarded and shape is distorted. As discussed previously, dichotomizing or trichotomizing continuous data invokes its assumptions of equality within groups and difference between groups *regardless* of the shape of the original data. X3 in Table 2, for example, is somewhat normally distributed. This is illustrated graphically in Figure 3. The conversion of the data in X3' substantially alters this distribution as illustrated in Figure 4. The results of a regression on X3 and an ANOVA on X3' are presented in Figure 5. As seen in the results of the regression, the results are statistically significant ($p_{\text{CALCULATED}} = .0001$) with an R^2 of .555. The results

from the ANOVA on the modified data ($X3'$), however, are not statistically significant ($p_{\text{CALCULATED}} = .115$) and the effect size, measured by η^2 , dropped to .214. Clearly, this is evidence that "statistical tests based upon continuous distributions are more powerful than those based upon dichotomized (or trichotomized) distributions" (Humphreys & Fleishman, 1974, p. 468, italics added). In this scenario, the re-expression of the original data has discarded a substantial amount of variance, which can mean "nonsignificant results when in fact the tested results may be significant" (Kerlinger & Pedhazur, 1973, p. 8).

A strategy sometimes resorted to *after* trichotomizing data is to drop the middle group and run analyses only with the extreme "high" and "low" groups (Humphreys, 1978). If this is done on the data in $X3'$ by eliminating the subjects with scores of 2, we obtain the results illustrated in Figure 6. Statistical significance is reached ($p_{\text{CALCULATED}} = .05$) and η^2 becomes .283. While these results appear to be "better" than the original ANOVA results, they are still lower than those produced by regression on the original $X3$ data. These results support the explanation of Humphreys (1978) regarding the elimination of middle groups, "While power can be increased after categorizing by discarding a category or categories from the center of the distribution, the increase does not compensate for the initial loss" (p. 873).

Humphreys (1978) goes on to explain another critical effect of eliminating middle groups, "as a side effect, the inflation of the differences between means when extreme groups are used frequently inflates the experimenter's evaluation of the importance of those differences" (p. 873). In other words, the experimenter is only looking at the extreme values from the trichotomized data. And, this occurs *after* converting the

original continuous data to nominal, trichotomized data! Without considering the scores in the middle, there may be a perceived importance of the variable as there is a large difference between the extremely low and extremely high scores. Again, in this case a regression on the original data would be more appropriate as it would prevent the loss of information due to eliminating or "mutilating" the original data.

Conclusion

The loss of variance that occurs when continuous or interval data is converted to nominal data is a major concern when considering the use of ANOVA as an analytic tool. This paper has reviewed specific concerns that should be considered prior to using OVA techniques on continuous predictor variables. The loss of information, reduction in reliability, and assumptions invoked by ANOVA are reasons why correlational techniques such as regression more fully preserve the nature of the original data. As Pedhazur (1982) emphasized:

Categorization of attribute variables is all too frequently resorted to in the social sciences.... It is possible that some of the conflicting evidence in the research literature of a given area may be attributed to the practice of categorization of continuous variables.... Categorization leads to a loss of information, and consequently to a less sensitive analysis. (pp. 452-453)

References

- Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt Brace Jovanovich.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.
- Cohen, J., & Cohen, P. (1983) Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Edgington, E.S. (1964). A tabulation of inferential statistics used in psychology journals. American Psychologist, 19, 202-203.
- Edgington, E.S. (1974). A new tabulation of statistical procedures used in APA journals. American Psychologist, 29, 25-26.
- Elmore, P.B., & Woehlke, P.L. (1988). Statistical methods employed in American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17(9), 19-20.
- Goodwin, L.D., & Goodwin, W.L. (1985). Statistical techniques in AERJ articles, 1979-1983: The preparation of graduate students to read the educational research literature. Educational Researcher, 14(2), 5-11.
- Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998) Applied statistics for the behavioral sciences. Boston: Houghton Mifflin Company.
- Humphreys, L.G. (1978). Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. Journal of Educational Psychology, 70, 873-876.

Humphreys, L. G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. Journal of Educational Psychology, 66, 464-472.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart and Winston.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd ed.). New York: Holt, Rinehart and Winston.

Song, M. X., & Parry, M. E. (1999). Challenges of managing the development of breakthrough products in Japan. Journal of Operations Management, 17, 665-688.

Stone-Romero, E. F., Weaver, A. E., & Glenar, J. L. (1995). Trends in research design and data analytic strategies in organizational research. Journal of Management, 21, 141-157.

Thompson, B. (1981). The problem of OVAism. Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, KY. (Order document #03980, National Auxillary Publication Service, P.O. Box 3513, Grand Central Station, New York, NY 10017)

Thompson, B. (1985). Alternate methods for analyzing data from experiments. Journal of Experimental Education, 54, 50-55.

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. Educational and Psychological Measurement, 46, 917-928.

Thompson, B. (1994). Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neo- classical perspective. In B. Thompson (Ed.), Advances in social science methodology (Vol. 3, pp. 3-27). Greenwich, CT: JAI Press.

Thompson, B. (1999, April). Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap. Invited address presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED 429 110)

Table 1

Breakdown of project scores

	<u>Number of Projects</u>		
	<u>0-2.5</u>	<u>2.5-7.5</u>	<u>7.5-10</u>
7-item scale	32	561	195

Note. From "Challenges of Managing the Development of Breakthrough Products in

Japan," by X. Michael Song and Mark E. Parry, 1999, *Journal of Operations*

Management, 17, p. 675.

Table 2

Heuristic Data Set

Id	Y	<u>Predictors</u>					
		X1	X1'	X2	X2'	X3	X3'
1	295	499	1	599	1	584	1
2	297	499	1	599	1	584	1
3	299	500	1	599	1	589	1
4	299	500	1	599	1	589	1
5	299	500	1	599	1	589	1
6	301	501	1	599	1	596	1
7	303	501	1	599	1	596	1
8	296	599	2	600	2	597	2
9	298	599	2	600	2	598	2
10	300	600	2	600	2	600	2
11	300	600	2	600	2	600	2
12	300	600	2	600	2	600	2
13	302	601	2	600	2	602	2
14	304	601	2	600	2	603	2
15	298	699	3	601	3	604	3
16	300	699	3	601	3	604	3
17	302	700	3	601	3	611	3
18	302	700	3	601	3	611	3
19	302	700	3	601	3	611	3
20	304	701	3	601	3	616	3
21	306	701	3	601	3	616	3

Note. X1', X2', and X3' are the re-expressions in nominal score form of their intervally-scaled variable counterparts.

Figure Captions

Figure 1. Distribution of scores for variable X1.

Figure 2. Distribution of scores for variable X1'.

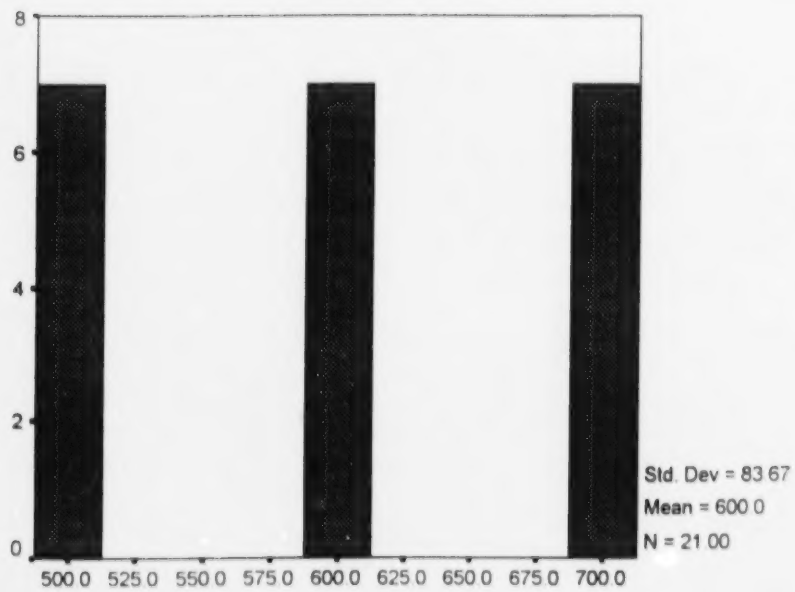
Figure 3. Distribution of scores for X3.

Figure 4. Distribution of scores for X3'.

Figure 5. Regression (Y, X3) and ANOVA (Y, X3') of Table 2 data.

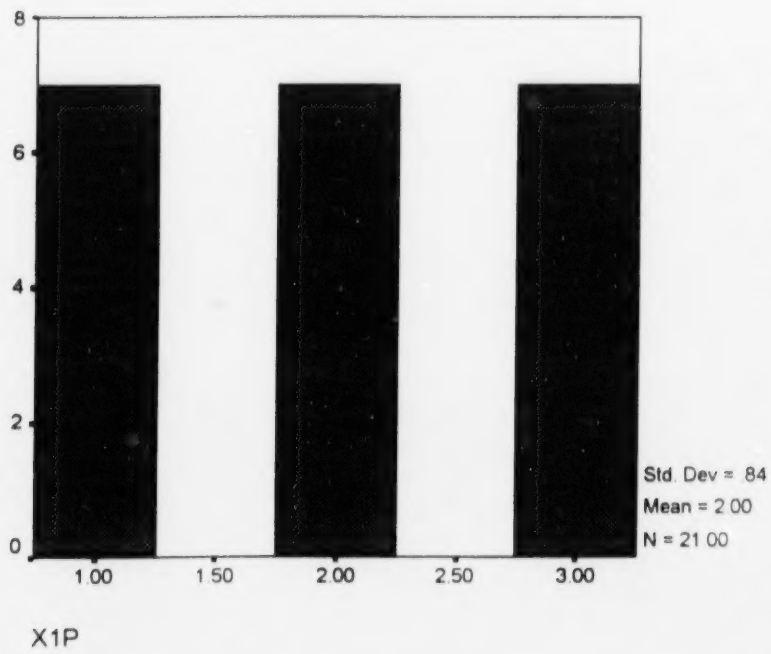
Figure 6. ANOVA (Y, X3') with middle group deleted.

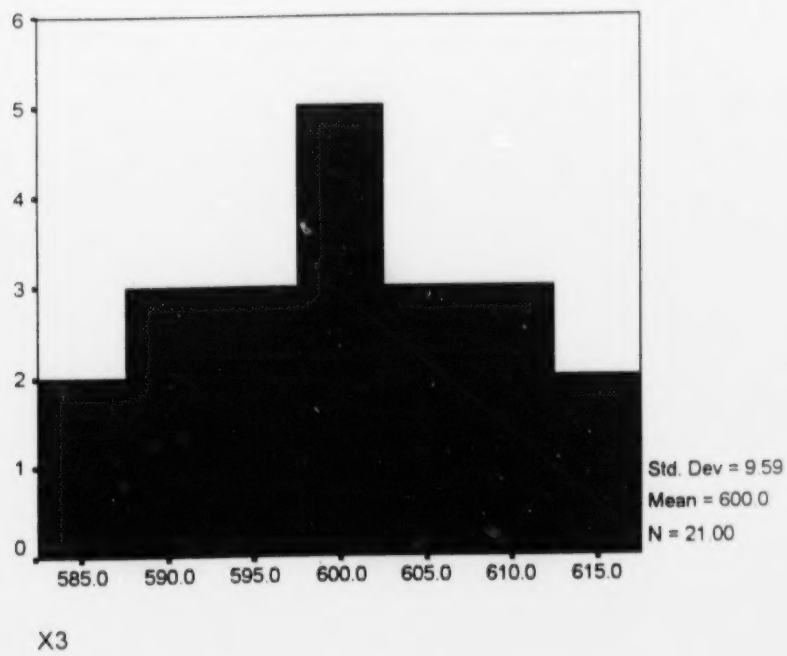
Dichotomizing Continuous Variables 18

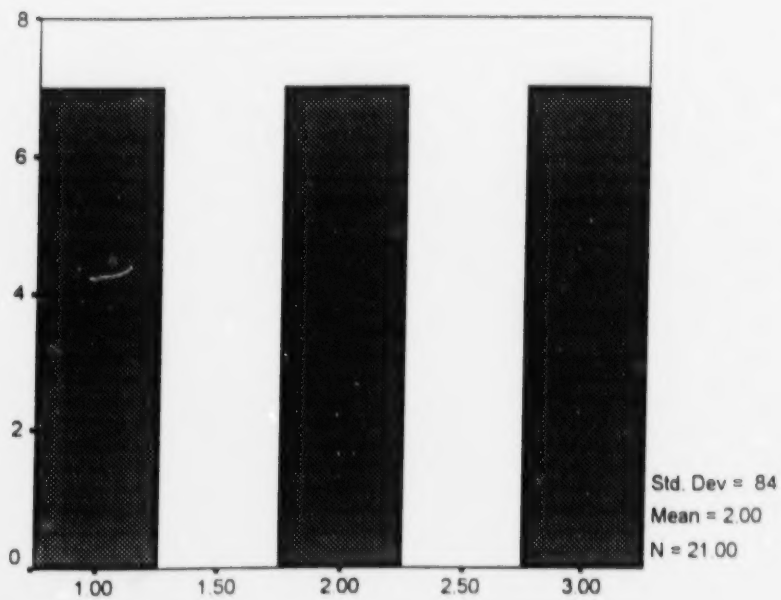


X1

Dichotomizing Continuous Variables 19







X3P

Regression (Y, X3)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.745 ^a	.555	.532	1.8900

a. Predictors: (Constant), X3

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	84.796	1	84.796	23.738	.000 ^a
	Residual	67.870	19	3.572		
	Total	152.667	20			

a. Predictors: (Constant), X3

b. Dependent Variable: Y

ANOVA (Y, X3')

ANOVA

Y					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	32.667	2	16.333	2.450	.115
Within Groups	120.000	18	6.667		
Total	152.667	20			

Note. Using an Excel function (i.e., " $\text{=FDIST}(f, df1, df2)$ " = " $\text{=FDIST}(23.738, 1, 19)$ "), the exact $p_{\text{CALCULATED}}$ value was found to be .000106. For the ANOVA, η^2 was computed to be 21.39% ($32.6667 / 152.6667$).

ANOVA

Y

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	31.500	1	31.500	4.725	.050
Within Groups	80.000	12	6.667		
Total	111.500	13			

Note For the ANOVA, η^2 was computed to be 28.25% ($31.50 / 111.50$).

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

**University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions**

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

**ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598**

Telephone: 301-497-4080

Toll Free: 800-799-3742

FAX: 301-953-0263

e-mail: ericfac@inet.ed.gov

WWW: <http://ericfac.piccard.csc.com>